GuardianNet

Công cụ bảo vệ trực tuyến bằng AI

Đội AIGenerated

GDGOC Hackathon Vietnam 2025

Thành viên: Nguyễn Hoàn Thiện (Đội trưởng)
Lê Kim Hoàng Trung
Nguyễn Lê Tất Phú
Doãn Cát Phú

12 tháng 4 năm 2025

Mục lục

1 Giới thiệu dự án

GuardianNet là giải pháp toàn diện nhằm bảo vệ người dùng Internet khỏi những nội dung không phù hợp như hình ảnh bạo lực, khiêu dâm và ngôn từ thô tục. Sản phẩm được phát triển với tiêu chí Responsible AI, nhắm đến việc tạo ra môi trường trực tuyến an toàn, đặc biệt cho trẻ em giúp tạo ra một môi trường Internet an toàn và lành mạnh hơn. Áp dụng công nghệ AI tiên tiến, đặc biệt là phương pháp multi-task learning, cho phép xây dựng mô hình nhận diện chính xác và hiệu quả, đồng thời tích hợp liền mạch vào trình duyệt thông qua extension.

Dựa trên những tiêu chí của một sản phẩm mang tính "Responsible AI", GuardianNet đảm bảo:

- **An toàn cho người dùng Internet:** Mục tiêu chính là bảo vệ người dùng, đặc biệt là trẻ em, khỏi tác động tiêu cực của nội dung độc hại, góp phần vào một không gian mạng lành mạnh hơn.
- **Tính công bằng:** Bất kỳ lứa tuổi nào khi truy cập vào Internet, những nội dung độc hại, phản cảm sẽ bị loại bỏ.
- **Tính minh bạch:** Cung cấp cho người dùng (ví dụ: các bậc phụ huynh) khả năng xem lại các nội dung đã bị chặn (nếu cần và đảm bảo riêng tư) và hiểu lý do tại sao chúng bị chặn. Người dùng có thể dễ dàng dùng như một extension trên trình duyệt của mình.
- **Bảo vệ quyền riêng tư và dữ liệu cá nhân của người dùng:** Thông tin được lấy trực tiếp từ trang web để ngăn chặn thông tin độc hại và phản cảm.

2 Vấn đề và phân tích

2.1 Xác định vấn đề

Mô tả vấn đề:

Có thể thấy rằng hiện nay trẻ em và người dùng dễ bị tiếp xúc với nội dung bạo lực và không phù hợp trên Internet, rất nhiều hình ảnh nhạy cảm, 18+, máu me vẫn có thể được truy cập và xem mà không có một biện pháp ngăn chặn nào. Các hệ thống lọc hiện có dựa trên từ khóa hay các tiêu chí cứng nhắc không thể phân biệt đầy đủ các trường hợp phức tạp, dẫn đến cả lỗi bỏ sót lẫn cảnh báo sai. Những hình ảnh đó khi bị cấm cũng dễ dàng xem được bằng nhiều cách khác, những hình ảnh máu me, bạo lực vẫn còn bỏ sót. Tạo nên một môi trường mạng không trong sạch, đặc biệt là với trẻ nhỏ.

2.2 Phân tích vấn đề

Có thể thấy thế giới Internet rất rộng lớn, việc dễ dàng tiếp cận với những nội dung không mấy lành mạnh ngay từ khi còn nhỏ có thể dẫn đến những hậu quả nghiêm trọng về tâm sinh lý, nhận thức ảnh hưởng xấu đến quá trình học tập và phát triển. Với người đang gặp các vấn đề tiêu cực hoặc nhân cách bất ổn có thể mắc các vấn đề như trầm cảm, tăng hành vi bạo lực.

Như đã đề cập trước đó, những giải pháp hiện tại còn thiếu tính linh hoạt, vẫn chưa được tập trung giải quyết một cách kỹ lưỡng, không thể cung cấp bảo vệ toàn diện cho

các đối tượng nhạy cảm, đặc biệt là trẻ em. Do vậy vẫn còn đó những mối nguy về việc tiếp cận những nội dung bạo lực và nhạy cảm đó.

3 Giải pháp – Phương án ứng dụng Responsible AI

Dựa trên những tiêu chí của Responsible AI và vấn đề về việc quản lý nội dung trên Internet đối với trẻ nhỏ, GuardianNet với vai trò là tiện ích mở rộng trên trình duyệt, tích hợp mô hình AI tiên tiến với phương pháp multi-task learning nhằm tự động quét và xử lý hình ảnh chứa nội dung bạo lực. Bao gồm các đặc trưng như:

3.1 Quét và phân tích nội dung

Khi người dùng duyệt web, content script quét các hình ảnh trên trang và gửi dữ liệu về background script. Tại đây, mô hình multi-task learning được sử dụng để xác định:

- Hình ảnh có chứa nội dung bạo lực, máu me, nhạy cảm hay không. Nếu có, hình ảnh sẽ bị làm mờ đi hoặc chặn hiển thị.
- Sự xuất hiện của con người, giúp giảm cảnh báo sai từ các hình ảnh không liên quan (ví dụ: phong cảnh).

Ngoài ra, GuardianNet còn quét nội dung văn bản để phát hiện các từ ngữ thô tục, phản cảm, lăng mạ dựa trên danh sách từ khóa và ngữ cảnh câu. Các từ/cụm từ này có thể được ẩn đi hoặc thay thế bằng ký tự [***].

3.2 Xử lý nội dung tự động

Khi extension được kích hoạt, những hình ảnh trên web sẽ tự động được load và đưa vào pre-trained model MobileNetV2 kết hợp với mô hình đa tác vụ giúp nhận diện ảnh bạo lực và sự hiện diện của người. Các kết quả của hai nhiệm vụ được mã hóa thành chuỗi nhị phân "ab". Với quy tắc:

- "a" biểu thị kết quả của nhiệm vụ bạo lực (0: không bạo lực, 1: bạo lực).
- "b" biểu thị kết quả của nhiệm vụ phát hiện con người (0: không có người, 1: có người).

Hình ảnh sẽ được quyết định có làm mờ hay không dựa vào mã nhị phân "ab", đảm bảo những ảnh an toàn được hiển thị:

- "11": Bạo lực + có người → Làm mờ.
- "10": Bạo lực + không người → Làm mờ.
- "01": Không bạo lực + không người → Không làm mờ.

3.3 Đảm bảo quyền riêng tư và dữ liệu cá nhân của người dùng

Hầu hết thông tin sẽ được lấy trực tiếp từ web (bao gồm hình ảnh, văn bản) khi người dùng sử dụng môi trường mạng để có thể ngăn chặn những thông tin tiêu cực và nhạy cảm đó.

3.4 Báo cáo minh bạch

Hệ thống ghi lại các hoạt động xử lý nội dung, cung cấp báo cáo định kỳ cho phụ huynh và quản trị viên. Ngoài ra các bậc phụ huynh có thể xem và hiểu lý do tại sao những nội dung đó lại bị chặn, hạn chế con em mình tiếp cận với những nội dung đó một lần nữa.

4 Chi tiết sản phẩm

4.1 Tính năng cốt lõi

4.1.1 Nhận diện nội dung bạo lực dưới dạng hình ảnh hoặc văn bản

Sử dụng mô hình multi-task learning với MobileNetv2 làm base model, hệ thống phân tích hình ảnh theo thời gian thực với hai nhiệm vụ:

- Phát hiện nội dung bạo lực.
- Phát hiện sự xuất hiện của con người.

Kết quả từ hai nhiệm vụ được kết hợp (theo quy tắc nhị phân "ab") để xác định chính xác hình ảnh cần làm mờ.

4.1.2 Cấu hình cá nhân hóa

Cho phép người dùng tùy chỉnh mức độ nhạy cảm, thiết lập danh sách trắng/đen và cấu hình các tham số báo cáo, đáp ứng nhu cầu riêng của gia đình hoặc môi trường giáo dục.

4.1.3 Báo cáo và giám sát

Ghi lại và trình bày báo cáo chi tiết về các hình ảnh đã được xử lý, cung cấp thông tin về thời gian xử lý, số lượng hình ảnh và các thống kê liên quan. Dựa vào đó, phụ huynh có thể xem lý do vì sao những hình ảnh hay text bị chặn để có thể đưa ra biện pháp xử lý.

4.1.4 Tích hợp liền mạch

Extension được tích hợp vào các trình duyệt phổ biến (Chrome, Edge, Brave) với giao diện người dùng trực quan và dễ sử dụng, đảm bảo quá trình duyệt web không bị gián đoạn.

4.2 Giao diện người dùng & Tài liệu hướng dẫn sử dụng

4.2.1 Giao diện

Giao diện người dùng của GuardianNet được thiết kế đơn giản, trực quan với các thành phần:

- Pop-up hiển thị trạng thái của bộ lọc.
- Trang Options cho phép điều chỉnh cấu hình và xem báo cáo.

4.2.2 Hướng dẫn sử dụng

Các bước cài đặt extension:

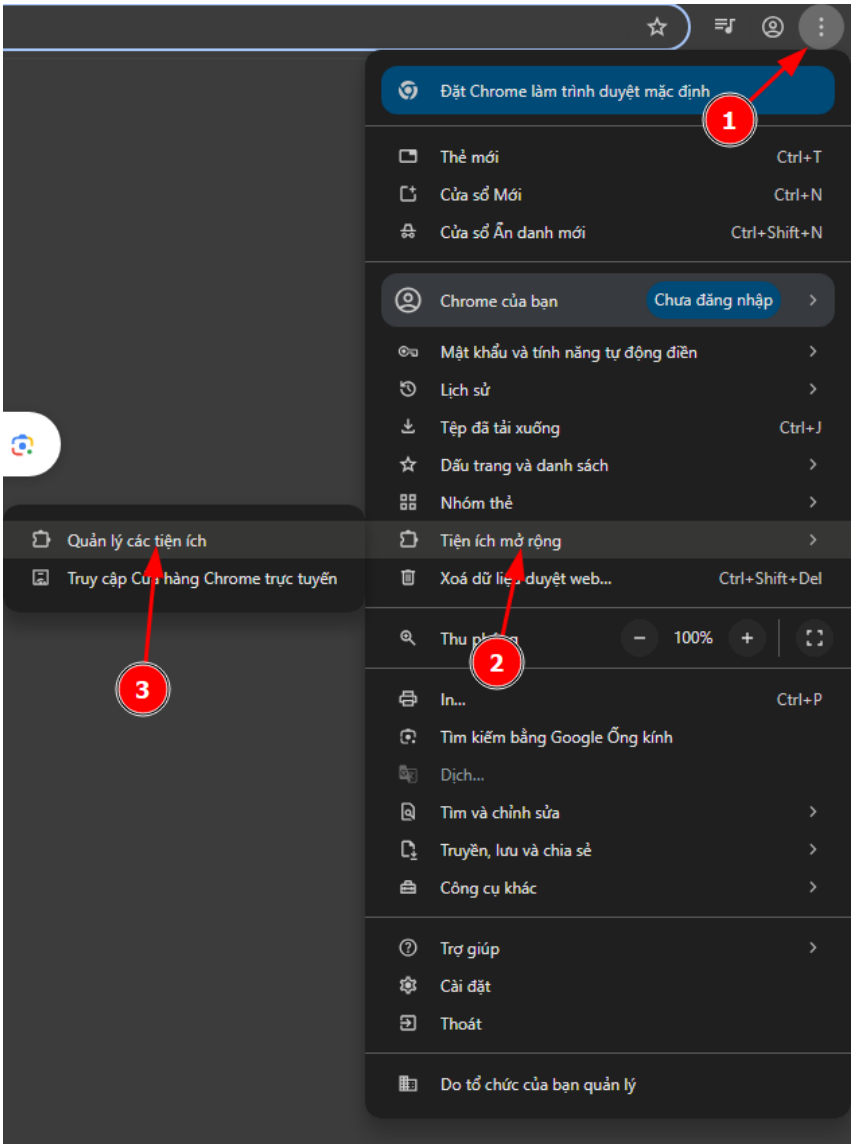
Bước 1: Chuẩn bị extension

```
# 1. Clone repository
git clone [repository-url]

# 2. Install dependencies
npm install

# 3. Build extension
npm run build
```

Hoặc tải file zip chúng tôi cung cấp: link [], giải nén và thực hiện các bước #4

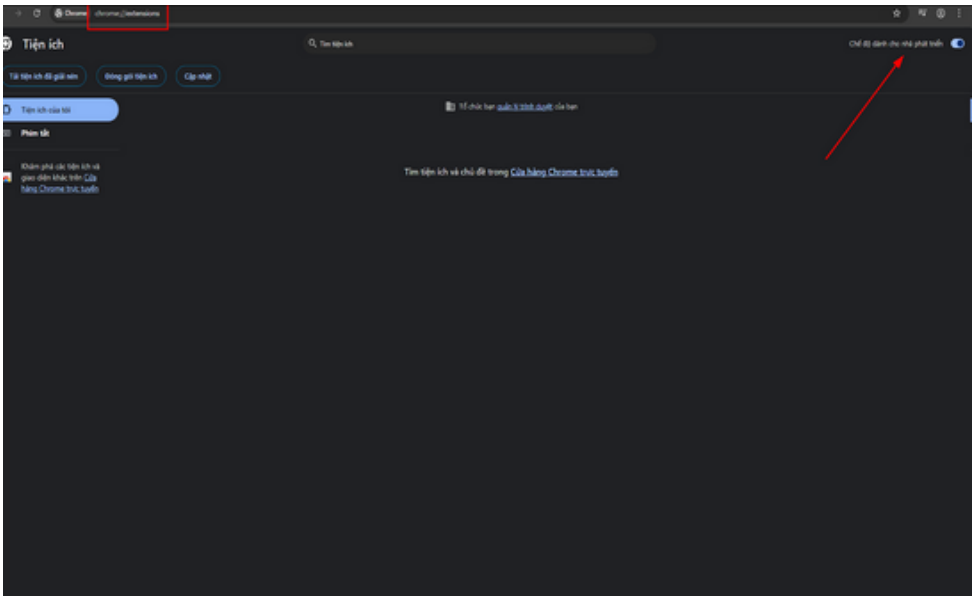
Bước 2: Load extension vào Chrome:

1. Truy cập chrome://extensions/



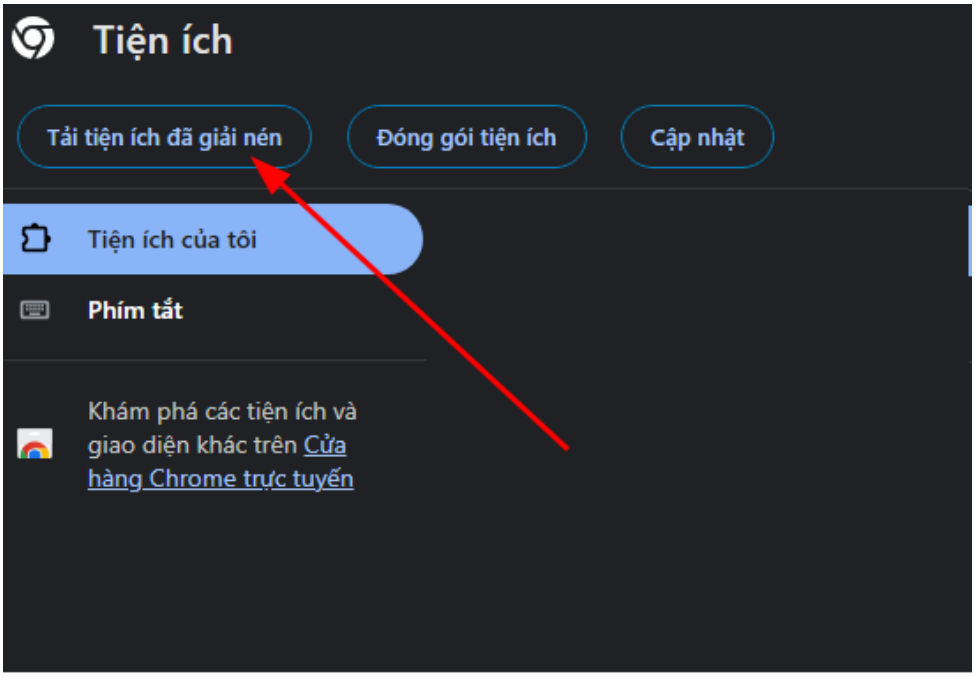
Hình 1: Trang Extensions của Chrome

2. Bật "Chế độ nhà phát triển"



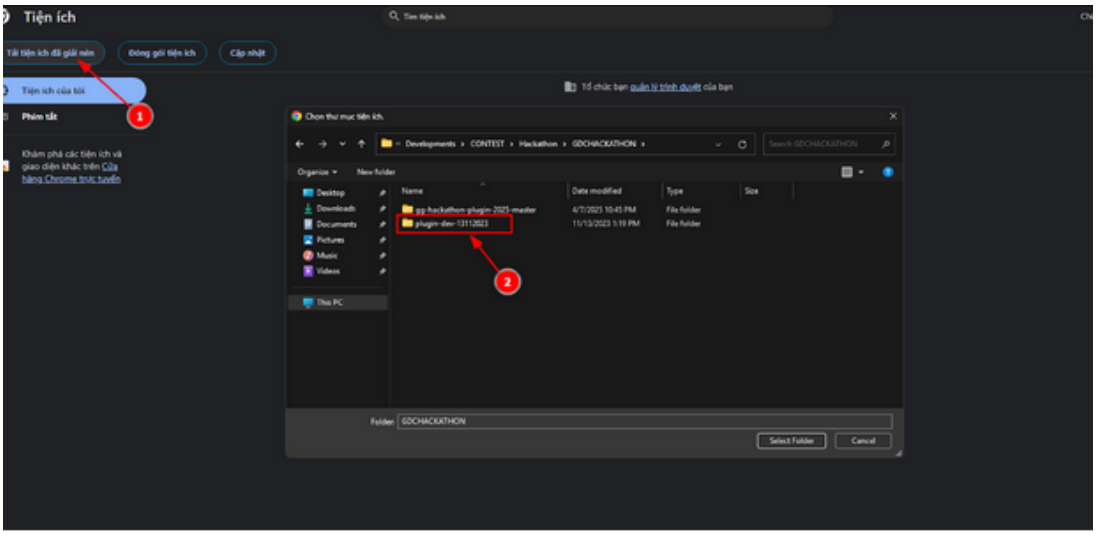
Hình 2: Kích hoạt chế độ nhà phát triển

3. Click "Tải tiện ích đã giải nén"



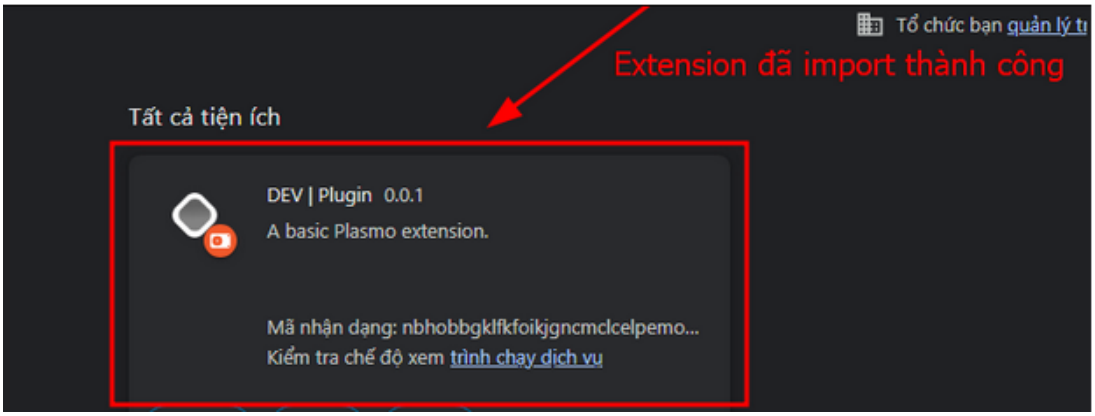
Hình 3: Tải extension đã giải nén

4. Chọn thư mục build của project



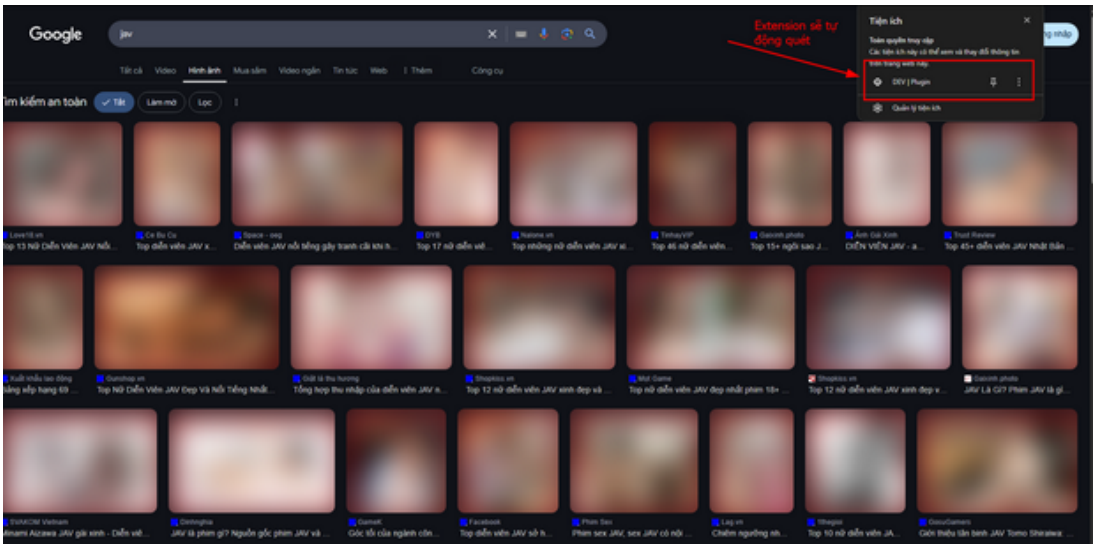
Hình 4: Chọn thư mục build

5. Extension đã load thành công!



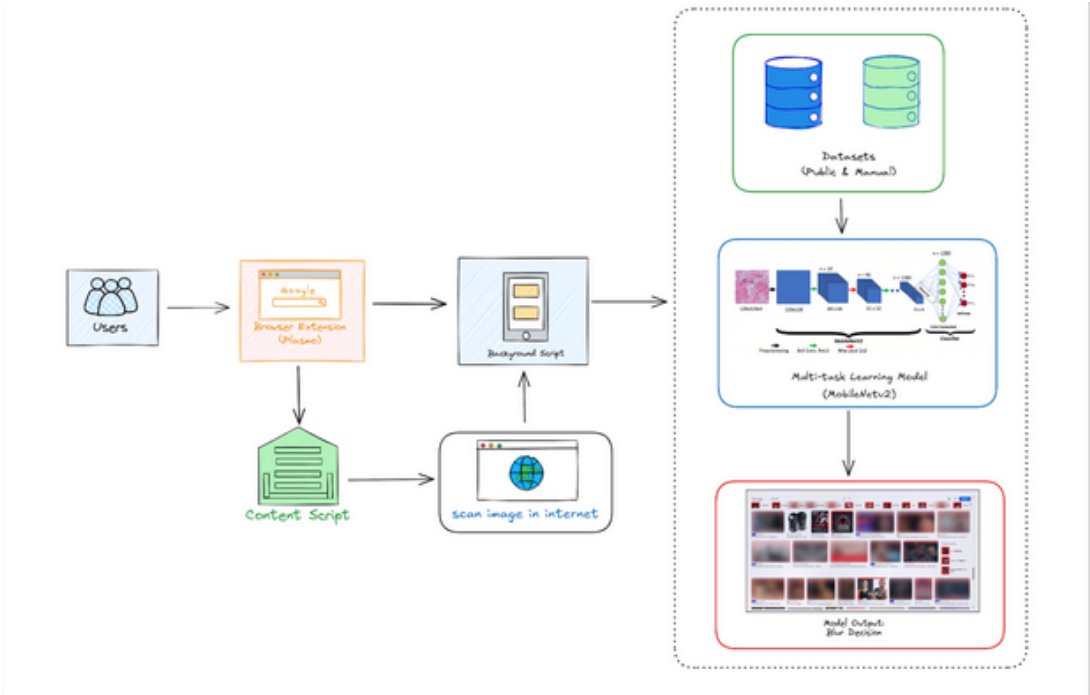
Hình 5: Extension đã cài đặt thành công

Giờ bạn có thể bật extension và test thử bằng cách truy cập những trang chứa nội dung nhạy cảm, bạo lực,...



Hình 6: Giao diện extension khi kích hoạt

5 Biểu đồ kiến trúc và công nghệ



Hình 7: Kiến trúc tổng thể của GuardianNet

5.1 Kiến trúc hệ thống & Mô hình AI

5.1.1 Multi-task Learning Model

Học đa tác vụ giúp huấn luyện model học được nhiều tác vụ liên quan với nhau cùng lúc, không như model đơn tác vụ chỉ có thể học một tác vụ duy nhất. Bằng cách cho model

học đa tác vụ, nó có thể học những đặc trưng chung có từ nhiều tác vụ liên quan, giúp tăng cường nhận diện hình ảnh bạo lực. Với 2 tác vụ:

- Phát hiện nội dung bạo lực.
- Phát hiện sự xuất hiện của con người.

5.1.2 Base Model

Mô hình học đa tác vụ gồm base model là MobileNetv2 sử dụng các phép tích chập có thể tách theo chiều sâu để xây dựng các mô hình nhẹ với chi phí tính toán giảm so với các phép tích chập tiêu chuẩn. MobileNetv2 đã được huấn luyện trước trên ImageNet, đảm bảo kích thước nhỏ và hiệu năng cao.

5.1.3 Các tác vụ được huấn luyện

Gồm có 2 tác vụ:

- Phát hiện hình ảnh mang nội dung bạo lực.
- Phát hiện sự xuất hiện của con người.

5.1.4 Kết hợp kết quả

Sử dụng quy tắc nhị phân "ab" để đưa ra quyết định cuối cùng: làm mờ nếu nhận nhãn "11" (có bạo lực và có sự xuất hiện của con người).

5.2 Tích hợp vào Web Extension

5.2.1 Framework Plasmo

Extension được xây dựng trên nền tảng Plasmo, một framework hiện đại giúp phát triển các tiện ích mở rộng cho trình duyệt một cách nhanh chóng, dễ bảo trì và tối ưu hiệu năng.

5.2.2 Cấu trúc Extension

- **Manifest V3:** Định nghĩa quyền truy cập, background script và content script.
- **Background Script:** Quản lý tải mô hình và xử lý yêu cầu từ content script.
- **Content Script:** Quét các hình ảnh trên trang, gửi dữ liệu về background và áp dụng hiệu ứng làm mờ khi cần.

5.3 Công nghệ và công cụ hỗ trợ

5.3.1 Ngôn ngữ & Frameworks

- **Python:** Dùng để huấn luyện mô hình AI với TensorFlow và Keras.
- **MobileNetv2:** Là kiến trúc chính cho mô hình đa tác vụ.
- **JavaScript, HTML5, CSS3:** Sử dụng cùng với Plasmo để xây dựng giao diện và tích hợp extension.

5.3.2 API & Dịch vụ

- **Chrome Extensions API:** Để tích hợp mô hình và xử lý hình ảnh trực tiếp trên trình duyệt.

5.3.3 Datasets

Kết hợp dữ liệu từ các bộ dữ liệu công khai (Real Life Violence Situations Dataset, HAR Dataset) và dữ liệu thu thập thủ công từ các nền tảng xã hội, đảm bảo tính đa dạng và thực tế trong huấn luyện mô hình.

6 Các công cụ, Dependencies, Frameworks và Datasets

6.1 Dependencies và Frameworks

- **Front-End:** JavaScript, HTML5, CSS3. Framework Plasmo để phát triển extension.
- **AI/ML:** Python, TensorFlow, Keras, MobileNetv2.

6.2 Datasets

- **Hình ảnh:** Dữ liệu công khai kết hợp với dữ liệu thu thập thủ công từ các nguồn xã hội.
- **Văn bản & Danh sách từ khóa:** Hỗ trợ việc xử lý ngữ cảnh và tăng độ chính xác nhận diện.

6.3 Công cụ hỗ trợ

- **API:** Chrome Extensions API
- **Quản lý dự án:** Git, GitHub, và các công cụ hỗ trợ phát triển khác.

7 Kế hoạch phát triển & Tiềm năng mở rộng

7.1 Kế hoạch phát triển

7.1.1 Giai đoạn 1 (Hackathon)

- Phát triển phiên bản MVP với chức năng quét và xử lý hình ảnh cơ bản thông qua mô hình đa tác vụ.
- Đảm bảo giao diện và hướng dẫn sử dụng thân thiện, dễ tiếp cận cho người dùng thử nghiệm.

7.1.2 Giai đoạn 2 (Sau Hackathon)

- Nâng cao độ chính xác của mô hình AI, bổ sung các chức năng báo cáo chi tiết và cấu hình nâng cao.
- Tích hợp thêm các cơ chế bảo mật và quản trị người dùng nhằm tăng cường tính minh bạch.

7.1.3 Giai đoạn 3 (Phát triển dài hạn)

- Mở rộng ứng dụng sang nhiều nền tảng (mobile, desktop) và đối tượng sử dụng (trường học, doanh nghiệp, gia đình).
- Nghiên cứu tích hợp thêm các công nghệ như deepfake detection, nhận diện ngữ cảnh nâng cao để mở rộng phạm vi bảo vệ.

7.2 Tiềm năng mở rộng

Ứng dụng đa dạng: GuardianNet có thể được áp dụng trong các môi trường gia đình, trường học và các tổ chức nhằm bảo vệ trẻ em và người dùng khỏi nội dung không phù hợp.

Khả năng thị trường: Với xu hướng ngày càng tăng của việc tiêu thụ nội dung trên Internet và nhu cầu bảo vệ người dùng, sản phẩm có tiềm năng mở rộng ra nhiều quốc gia và lĩnh vực khác nhau.

8 Kết luận

GuardianNet là giải pháp toàn diện ứng dụng công nghệ AI tiên tiến kết hợp phương pháp multi-task learning để nhận diện và lọc bỏ nội dung bạo lực trên Internet. Sản phẩm không chỉ đạt độ chính xác cao (98.5%) mà còn tích hợp liền mạch vào trình duyệt, giúp tạo ra môi trường duyệt web an toàn và thân thiện cho trẻ em. Với tiêu chí Responsible AI, GuardianNet cam kết về sự công bằng, minh bạch, bảo mật và khả năng mở rộng trong tương lai.

9 Thông tin liên hệ

Thông tin liên hệ đội AIGenerated

Nguyễn Hoàn Thiện (Leader)	Email: thiennh.21it@vku.udn.vn SĐT: 0356496977
Lê Kim Hoàng Trung (Member)	Email: trunglkh.21it@vku.udn.vn SĐT: 0962043095
Nguyễn Lê Tất Phú (Member)	Email: phunlt.21it@vku.udn.vn SĐT: 0522944603
Doãn Cát Phú (Member)	Email: phudc.21it@vku.udn.vn SĐT: 0935026145